Preface

NEEDS AND REQUIREMENTS FOR INFORMATION RETRIEVAL

Scientific and economic organizations are confronted with handling an abundance of strategic information in their domain activities. One main challenge is to be able to find the right information quickly and accurately. In order to do so, organizations must master information access: getting relevant query results that are organized, sorted, and actionable.

As noted by Mukhopadhyay and Mukhopadhyay (2004), almost everyone agrees that in the current state of the art on Internet search engine technology, extracting information from the Web is an art itself. Almost all commercial search engines use classical keyword-based methods for information retrieval (IR). That means that they try to match user specified patterns (i.e., queries) to the texts of all documents in their database and then return the documents that contain terms matching the query. Such methods are quite effective for well-controlled collections - such as bibliographic CD-ROMs or handcrafted scientific information repositories. Unfortunately the organization of the Internet has not been rationally supervised, but it has rather spontaneously evolved and, therefore, cannot be treated as a well-controlled collection. It contains a lot of garbage and redundant information and, what is maybe even more important, it does not rely on any underlying semantic structure intended to facilitate navigation.

In addition, some of the current issues result from inappropriate query constructions. The user queries that are usually submitted to search engines are often too general (like “water sources” or “capitals”) and this produces millions of returned documents. The results, which are of interest to users, are probably among them, but they cannot be distinguished from the mass; it appears impossible to emphasize them to the human attention. One hundred documents are generally regarded as the maximum amount of information that can be useful to users in such situations.

On the other hand, some documents cannot be retrieved because the specified pattern does not exactly match. This can be caused by flexion in some languages, or by confusion introduced by synonyms and complex idiom structures (e.g., in English the word Mike is often given as an example of this, as it can be used as a male name or as a shortened form for the noun “microphone”). Most search engines have also very poor user interfaces. Computer-aided query constructions are very rare and the presentation of the search results concentrates mostly on individual documents, but it does not provide any general overview of retrieved data, which is crucial when the number of returned documents is huge. A last group of problems comes from the nature of information stored on the Internet. Search tools must not only deal with hypertext documents (in the form of WWW pages) but also with text repositories (message archives, e-books etc.), FTP and Usenet servers and with many sources of non-textual information such as audio, video, and interactive contents.

Recent technological progress in computer science, Web technologies, and constantly evolving information available on the Internet has drastically changed the landscape of search and access to information. Web search has significantly evolved in recent years. In the beginning, web search engines such as Google and Yahoo! were only providing search service over text documents. Aggregated search was one of the first steps to go beyond text search, and was the beginning of a new era for information seeking and retrieval. These days, new web search engines support aggregated search over a number of vertices, and blend different types of documents (e.g., images, videos) in their search results. New search engines employ advanced techniques involving machine learning, computational linguistics and psychology, user interaction and modeling, information visualization, Web engineering, artificial intelligence, distributed systems, social networks, statistical analysis, semantic analysis, and technologies over query sessions.

Documents no longer exist on their own; they are connected to other documents, they are associated with users and their position in a social network, and they can be mapped onto a variety of ontologies. Similarly, retrieval tasks have become more interactive and are solidly embedded in a user's geospatial, social, and historical context. It is conjectured that new breakthroughs in information retrieval will not come from smarter algorithms that better exploit existing information sources, but from new retrieval algorithms that can intelligently use and combine new sources of contextual metadata.

With the rapid growth of web-based applications, such as search engines, Facebook, and Twitter, the development of effective and personalized information retrieval techniques and of user interfaces is essential. The amount of shared information and of social networks has also considerably grown, requiring metadata for new sources of information, like Wikipedia and ODP. These metadata have to provide classification information for a wide range of topics, as well as for social networking sites like Twitter, and Facebook, each of which provides additional preferences, tagging information and social contexts. Due to the explosion of social networks and other metadata sources, it is an opportune time to identify ways to exploit such metadata in IR tasks such as user modeling, query understanding, and personalization, to name a few. Although the use of traditional metadata such as html text, web page titles, and anchor text is fairly well-understood, the use of category information, user behavior data, and geographical information is just beginning to be studied.

OBJECTIVES OF THE BOOK

The main goal of this book is to transfer new research results from the fields of advanced computer sciences and information science to the design of new search engines. The readers will have a better idea of the new trends in applied research. The achievement of relevant, organized, sorted, and workable answers – to name but a few – from a search is becoming a daily need for enterprises and organizations, and, to a greater extent, for anyone. It does not consist of getting access to structural information as in standard databases; nor does it consist of searching information strictly by way of a combination of key words. It goes far beyond that. Whatever its modality, the information sought should be identified by the topics it contains, that is to say by its textual, audio, video or graphical contents. This is not a new issue. However, recent technological advances have completely changed the techniques being used. New Web technologies, the emergence of Intranet systems and the abundance of information on the Internet have created the need for efficient search and information access tools.

TARGET AUDIENCE

This book is intended for scientists and decision-makers who wish to gain working knowledge of searches in order to evaluate available solutions and to dialogue with software and data providers. It also targets intranet or Web server designers, developers and administrators who wish to understand how to integrate search technology into their applications according to their needs. This book is further designed for designers, developers and administrators of databases, groupware applications and document management systems (EDM), as well as directors of libraries or documentation centers who seek a deeper understanding of the tools they use, and how to set up new information systems. Lastly, this book is aimed at all professionals in technology or competitive intelligence and, more generally, the specialists of the information market.

A BRIEF OVERVIEW OF THE ORGANIZATION OF THE BOOK

The book is divided into four sections:

Section 1 is "Indexation". The goal of automatic indexing is to establish an index for a set of documents that has to facilitate future access to documents and to their content. Usually, an index is composed of a list of descriptors, each of them being associated to a list of documents and/or of parts of documents to which it refers. In addition, theses references may be weighted. When searching to answer the users' queries, the system looks for a list of answers, of which an index is as close as possible to the demand. As a consequence, indexation could be seen as a required preliminary to intelligent information retrieval, since it pre-structures textual data according to topic, domain, keyword or center of interest.

Section 2 is "Data Mining for Information Retrieval". Data Mining (i.e., Knowledge Discovery from Data Bases) is the process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible patterns from large data sets. Data mining is a relatively young and interdisciplinary field that combines methods from statistics and artificial intelligence with database management. With the considerable increase of processing power, storage capacities, and inter-connectivity of computer technology, in particular with the grid computation, data mining is now seen as an increasingly important field by modern business for transforming unprecedented quantities of digital data into new knowledge that provides a significant competitive advantage. This is now a large part of what people refer to as business intelligence strategy. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The growing consensus that data mining can bring real added value has led to an explosion in demand for novel data mining technologies.

Section 3 is "Interface". The term "interface" refers to the part of the search engine in which (1) the user formulates his request and (2) the user reads the results. The interface is then seen in four views: Human-centered Web Search, Personalization, Question/Answering, and Mobile Search Engines. "Human-centered Web Search" is understood to be how Web search engines help people to find the information they are seeking. "Personalization" takes keywords from the user as an expression of their information need, but also uses additional information about the user (such as their preferences, community, location or history) to assist in determining the relevance of pages. "Question/Answering" addresses the problem of finding answers to questions posed in natural language; answering is the task which, when given a query in natural language, aims at finding one or more concise answers in the form of sentences or phrases. "Mobile Search Engines" may be defined as the combining of search technologies and knowl-

edge about the user context in his mobile environment into a single framework in order to provide the most appropriate answer for users information needs.

Finally, Section 4 is "Evaluation". Evaluation means two things: (1) tracing the users' behaviors, with a special attention to the concept of "information practice" and other related concepts such as "use", "activity", and "behavior" largely used in the literature but not always strictly defined, the aim being to place the users and their needs at the center of the design process; (2) evaluating the next generation search engines with four main criteria for improving the quality of the search results: index quality, quality of the results, quality of search features, and search engine usability.

Christophe Jouis
University Paris Sorbonne Nouvelle and LIP6 (UPMC & CNRS), France

Ismaïl Biskri
University of Quebec at Trois Rivieres, Canada

Jean-Gabriel Ganascia
LIP6, (UPMC & CNRS), France

Magali Roux
INIST and LIP6, (UPMC & CNRS), France

REFERENCE

Mukhopadhyay, B., & Mukhopadhyay, S. (2004, February 11-13). Data mining techniques for information retrieval. In *Proceedings of the 2nd International Conference of the Convention on Automation of Libraries in Education and Research Institution*, New Delhi, India (p. 506).